Outlier detection and trimmed estimation in general functional spaces

Daniel Gervini
Department of Mathematical Sciences
University of Wisconsin–Milwaukee
gervini@uwm.edu

January 7, 2010

Abstract

This article introduces trimmed estimators for the mean and covariance functional of data in general Hilbert spaces. The estimators are based on a data depth measure that can be computed on any Hilbert space, because it is defined only in terms of the interdistances between data points. We show that the estimators can attain the maximum breakdown point by properly choosing the tuning parameters, and that they possess better outlier resistance properties than alternative estimators, as shown by a comparative Monte Carlo study. The data depth measure we introduce can also be used for visual screening of the data and is a practical tool to detect clusters and isolated outliers, as shown by three real-data applications.

Key Words: Abstract Inference; Data depth; Outlier detection; Robust statistics; Stochastic processes.

1 Introduction

Many statistical problems today involve the analysis of data that do not fit the classical univariate or multivariate frameworks. For example, the data sometimes is a sample of univariate curves, like growth curves, EEG curves, spectrograms, and periodograms. Mathematically we can think of these univariate curves as realizations of a stochastic process X in $L^2(\mathbb{R})$. The statistical analysis of this type of data has received much attention in recent years (see e.g. Ramsay and Silverman 2002, 2005, and references therein).

The space $L^2(\mathbb{R})$, being a separable Hilbert space, has many things in common with $\mathbb{R}^p$. Therefore, the statistical concepts are not substantially different. For example, the covariance operator and its eigenfunctions in $L^2(\mathbb{R})$ are a straightforward extension of the covariance matrix and its eigenvectors in $\mathbb{R}^p$, so principal component analysis is carried out in a virtually identical way. The crucial difference between $\mathbb{R}^p$ and $L^2(\mathbb{R})$ is the dimensionality. Concepts that are meaningful in a finite dimensional space, like volume or Mahalanobis distance, cannot be extended in a useful way to infinite dimensional spaces. This complicates the development of outlier-detection tools and the definition of equivariant robust estimators, because most multivariate robust estimators are either based on the Mahalanobis distance (Maronna 1976; Davies 1987) or in notions of volume and data depth based on simplices (Oja 1983; Liu 1990). Even in finite dimensional spaces, if the sample size is smaller than the dimension (the "$n < p$ problem", so common in Chemometrics and Microarray Data Analysis) traditional robust covariance estimators cannot be computed and alternatives must be sought (Filzmoser et al. 2008, Filzmoser et al. 2009, Hsieh and Hung 2009).

In this article we will introduce a measure of "outlyingness" (or data depth) that does not depend on notions of volume or Mahalanobis distance, and therefore can be used in general Hilbert spaces for the detection of atypical observations and the construction of robust estimators. We will focus on infinite dimensional spaces, but the estimators can also be used in finite dimensional spaces and are particularly useful when the "$n < p$ problem" is present.

But the first question that arises is whether one needs, in practice, sample

Figure 1: Handwritten Digits Example. Eight instances of the number "five".

spaces that are even more abstract than $L^2(\mathbb{R})$. The answer is yes. Consider, as a motivating example, the excitation–emission matrices used in Chemometrics (Mortensen and Bro 2006). Here the data is of the form $X_i(s,t)$, where X_i is the intensity of light emitted at wavelength t when the ith chemical compound is excited at wavelength s. The resulting phosphorescence surfaces constitute a sample in $L^2(\mathbb{R}^2)$. We cannot show this sample in a single plot, but we have created a movie that shows the 338 sample surfaces in quick succession (available as supplementary material). This movie helps visualize the salient modes of variability and also the fact that there are some outiers in the sample. More about this data will be said in Section 6. At this point, however, it is clear that an outlier-screening procedure cannot consist of simply watching a tiresome movie that shows every sample surface; a more sophisticated and practical method is needed.

As a second example, consider a handwritten digit recognition problem. The horizontal trajectory of the pen tip can be seen as a curve $(x(t), y(t))$ in $\mathbb{R}^2$, where x is the horizontal position, y is the vertical position, and t is time. The sample space here is $L^2(\mathbb{R}) \times L^2(\mathbb{R})$. A few digits are shown in Figure 1; they are all drawings of the digit "five", although some of them look like a "six". The reason is that there are two ways of drawing the number "five", as explained in Section 6. This creates

two different clusters in the sample of "fives", which are very difficult to determine without an outlier-detection tool (the sample contains 1055 handwritten "fives", so visual screening is out of the question).

The problem of robust estimation in infinite-dimensional spaces has been addressed by other authors, such as Locantore et al. (1999), Fraiman and Muniz (2001), Cuevas et al. (2007), Gervini (2008), López-Pintado and Romo (2009), and Cuevas and Fraiman (2009). The approach varies. Locantore et al. (1999) and Gervini (2008) define estimators based on the spatial median and the covariance of normalized observations; the main drawback of these estimators is the low breakdown point of the principal components. The other papers are mainly concerned with data depth definitions and robust estimation is mostly seen as a by-product. The robustness properties of the estimators are therefore not well studied, and some of these data depth measures (specifically, the ones in Fraiman and Muniz 2001 and López-Pintado and Romo 2009) are based on the ranks of the $X_i(t)$s and cannot be easily extended beyond $L^2(\mathbb{R})$. Strictly speaking, they do not even apply to the unrestricted space $L^2(\mathbb{R})$, since they assume that the X_is can be evaluated at each t. Formally, this problem can be circumvented by restricting the definition to the space of continuous square-integrable functions, but the real practical problem is to obtain accurate estimates of $X_i(t)$ for each t, especially when the X_is are sparsely observed.

In this paper we propose an outlyingness measure that is entirely based on the interdistances between pairs of observations. Since any Hilbert space has a canonical norm and therefore a canonical metric associated with its inner product, this outlyingness measure can be defined on any Hilbert space. In fact, this measure could be defined on any metric space, but notions of directions of variability and principal components only make sense in spaces with inner products, so we will focus on Hilbert spaces in this paper.

The idea of defining robust estimators based on interdistances is not entirely new. For example, Wang and Raftery (2002) used a nearest-neighbor cleaning approach to defined robust covariance estimators. But this approach, as presented by the authors, cannot be extended to infinite-dimensional spaces (their weighting scheme depends on distributional assumptions that are only valid in finite-

dimensional spaces). In contrast, the estimators we define are trimmed estimators based on the ranks of the outlyingness measures, so no distributional assumptions are needed.

These estimators are very easy to compute and show very good robustness properties in our simulations. In fact, we derive their breakdown point and show that they can attain the maximum 0.50 by appropriately choosing the tuning parameters. The definitions of the outlyingness measure and the trimmed estimators are presented in Section 2. Their finite-sample properties are studied in Section 3, and their population properties in Section 4. The comparative Monte Carlo study is reported in Section 5, and the two real-data applications mentioned before are presented in Section 6. The Appendix contains some proofs.

2 Trimmed mean and covariance estimators: definition

Given a sample $X_1, \ldots, X_n$ in a Hilbert space $\mathcal{H}$, consider the interdistances $d_{ij} = \|X_i - X_j\|$. An observation X_i is an outlier if it is far from *most* of the other observations (not necessarily from *all* of them, since outliers sometimes form clusters). Given $\alpha \in [0, .50]$ define the α-radius r_i as the radius of the smallest ball centered at X_i that contains $100\alpha\%$ of the observations. This is easy to compute: for each i, the interdistances $\{d_{ij}\}$ are sorted in the index j, obtaining the sequence $d_{i,(1)}, \ldots, d_{i,(n)}$; then $r_i = d_{i,(\lceil \alpha n \rceil)}$, where $\lceil x \rceil$ denotes the integer closest to x from above.

The idea behind this definition is that the r_is will be small where the data is tight (usually near the center of the distribution) and large where the data is sparse, including regions with outliers. However, if there is a tight cluster of n^* outliers and $\lceil \alpha n \rceil < n^*$, the r_is for the outliers will be small, perhaps smaller than for the "good" data. For this reason, it is important to chose α large enough that at least one good observation will be captured by r_i if X_i is an outlier. In general, only $\alpha = .50$ will guarantee this (see Proposition 2 in Section 3). In practice, however, it is instructive to draw box-plots and histograms of the r_is for different values of

α; the outliers tend to emerge clearly and consistently as α increases.

The next step is to cut off or downweight the observations with largest radii. Given a trimming proportion $\beta \in [0, .5]$, define $w(X_i) = \mathbb{I}\{r_i < r_{(\lceil (1-\beta)n \rceil)}\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function, or more generally $w(X_i) = g(\mathrm{rank}(r_i)/n)$, where $\mathrm{rank}(r_i)$ is the rank of r_i among the radii and $g : [0, 1] \to \mathbb{R}^+$ is a bounded, non-negative and non-increasing function such that $g(t) > 0$ for $t < 1 - \beta$ and $g(t) = 0$ for $t \geq 1 - \beta$. We will mostly use hard-rejection weights in this paper, but if a smoother downweighting scheme is preferred, one can choose a $\beta_1 > \beta$ and take, for instance,

$$g(r) = \begin{cases} 1, & 0 \leq r \leq a, \\ (r-b)\left[\frac{1}{(a-b)} + \frac{(r-a)\{2r-(a+b)\}}{(b-a)^3}\right], & a \leq r \leq b, \\ 0, & r \geq b, \end{cases} \tag{1}$$

with $a = 1 - \beta_1$ and $b = 1 - \beta$, which is a differentiable function that progressively downweights the largest $100\beta_1\%$ radii and cuts off the largest $100\beta\%$ radii completely (one can take, for instance, $\beta_1 = .50$ and a less drastic $\beta = .20$).

The trimmed mean estimator is then defined as

$$\hat{\mu} = \frac{1}{\sum_{i=1}^n w(X_i)} \sum_{i=1}^n w(X_i) X_i \tag{2}$$

and the trimmed covariance functional as

$$\widehat{\mathfrak{C}} = \frac{1}{\sum_{i=1}^n w(X_i)} \sum_{i=1}^n w(X_i)(X_i - \hat{\mu}) \otimes (X_i - \hat{\mu}) \tag{3}$$

(the tensor product $f \otimes g$ is defined as $(f \otimes g)(h) = \langle f, h \rangle g$ for $h \in \mathcal{H}$, and $(f \otimes g)(h_1, h_2) = \langle f, h_1 \rangle \langle g, h_2 \rangle$ for $h_1, h_2 \in \mathcal{H}$).

As with univariate trimmed estimators, the trimming proportion β determines the amount of outliers that $\hat{\mu}$ and $\widehat{\mathfrak{C}}$ can tolerate, as well as their efficiency (Maronna et al. 2006, chap. 2; Van der Vaart 1998, chap. 22). The robustness increases with β but the efficiency decreases, so in practice we recommend to choose β in a data-driven way. A histogram of the radii usually gives a good idea of the proportion of

outliers in the sample, and we recommend to choose β large enough to cut off all the outliers, but not larger.

An important computational advantage of these estimators is that the weights $w(X_i)$ depend on the data only through the interdistances d_{ij}, which can be computed entirely from the inner products $\langle X_i, X_j \rangle$. These inner products can be typically well approximated by numerical integrals based on the raw observations of the X_is (see Gervini 2008, theorem 1); a finer reconstruction of the X_is is usually not required. This is an important practical advantage because the X_is are often sampled on sparse grids, and reconstructing them on a finer grid may be problematic (usually involving smoothing).

The principal components of $\widehat{\mathfrak{C}}$ can also be computed using the inner products $\langle X_i, X_j \rangle$ only. The principal components of $\widehat{\mathfrak{C}}$ are the (non-null) functions $\phi \in \mathcal{H}$ such that $\widehat{\mathfrak{C}}\phi = \lambda\phi$ for some $\lambda \in \mathbb{C}$, in which case λ is called an eigenvalue and ϕ an eigenfunction. Since $\widehat{\mathfrak{C}}$ is a compact self-adjoint operator, the eigenvalues are real and non-negative, and the number of strictly positive eigenvalues is countable when $\mathcal{H}$ is separable; in addition, each eigenvalue has a finite multiplicity, and eigenfunctions corresponding to different eigenvalues are orthogonal (Gohberg et al. 2003, chap. IV). Then, without loss of generality, we can assume that the principal components $\{\hat{\phi}_k\}$ are orthonormal and ordered so that $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots > 0$. As shown in Gervini (2008) or Jolliffe (2002, section 3.5), the principal components are linear combinations of the observations. Specifically, if $\tilde{w}_i = w(X_i)/\sum_{i=1}^n w(X_i)$, then $\hat{\phi}_k = \sum_{i=1}^n (c_{ki}/l_k^{1/2})\tilde{w}_i^{1/2}(X_i - \hat{\mu})$ and $\hat{\lambda}_k = l_k$, where $\mathbf{c}_k$ is the kth unit-norm eigenvector of the matrix $\mathbf{G} \in \mathbb{R}^{n\times n}$ with elements $G_{ij} = \langle \tilde{w}_i^{1/2}(X_i - \hat{\mu}), \tilde{w}_j^{1/2}(X_j - \hat{\mu}) \rangle$, and l_k is the corresponding kth eigenvalue. Note that, after some algebra, the G_{ij}s can be expressed entirely in terms of the inner products $\langle X_i, X_j \rangle$ and the $\tilde{w}_i$s.

3 Finite sample properties: equivariance and breakdown point

The reason our trimming scheme is based on interdistances is that the resulting weights $w(X_i)$ are invariant under translation, re-scaling and unitary transformations. As a consequence, the estimators (2) and (3) satisfy the natural equivariance properties of location and scatter estimators. This is shown in Proposition 1 below.

Recall that a unitary operator (the equivalent of an orthogonal matrix in $\mathbb{R}^p$) is a $\mathfrak{U} : \mathcal{H} \to \mathcal{H}$ such that $\|\mathfrak{U}f\| = \|f\|$ for every $f \in \mathcal{H}$, or equivalently, such that $\mathfrak{U}^*\mathfrak{U} = \mathfrak{U}\mathfrak{U}^* = \mathfrak{I}$, where $\mathfrak{I}$ is the identity operator and $\mathfrak{U}^*$ is the adjoint of $\mathfrak{U}$ (for an operator $\mathfrak{B} : \mathcal{H}_1 \to \mathcal{H}_2$, the adjoint is the unique operator $\mathfrak{B}^* : \mathcal{H}_2 \to \mathcal{H}_1$ that satisfies $\langle f, \mathfrak{B}g\rangle = \langle \mathfrak{B}^* f, g\rangle$ for every $g \in \mathcal{H}_1$ and $f \in \mathcal{H}_2$; this would be the equivalent of the transpose in $\mathbb{R}^p$). A unitary operator of particular interest is the "change of basis" operator $\mathfrak{U} = \sum \psi_k \otimes \phi_k$, where $\{\phi_k\}$ and $\{\psi_k\}$ are two orthonormal systems in $\mathcal{H}$.

Proposition 1 *Let $X_1, \ldots, X_n$ be a sample in $\mathcal{H}$, $a \neq 0$ a scalar, b a point in $\mathcal{H}$, and $\mathfrak{U}$ a unitary operator. Let $\tilde{X}_i = a\mathfrak{U}X_i + b$, and denote by $\tilde{d}_{ij}$, $\tilde{r}_i$, $\tilde{\mu}$, $\widetilde{\mathfrak{C}}$, $\tilde{\lambda}_k$ and $\tilde{\phi}_k$ the interdistances, radii and estimators computed with the transformed data. Then $\tilde{d}_{ij} = |a|d_{ij}$, $\tilde{r}_i = |a|\, r_i$, $w(\tilde{X}_i) = w(X_i)$, $\tilde{\mu} = a\mathfrak{U}\hat{\mu} + b$, $\widetilde{\mathfrak{C}} = a^2\mathfrak{U}\widehat{\mathfrak{C}}\mathfrak{U}^*$, $\tilde{\lambda}_k = a^2\hat{\lambda}_k$, and $\tilde{\phi}_k = \mathfrak{U}\hat{\phi}_k$.*

The proof is given in the Appendix. The key property here is that the radii are invariant under data rotations and shifts, and equivariant under scale changes. Thus, the ranks of the radii are invariant under the three transformations, as would be expected of an "outlyingness" measure: if an observation is deemed an outlier for certain data configuration, it should still be seen as an outlier if the dataset is simply shifted, rotated or rescaled.

Let us now turn to the outlier-resistance properties of the trimmed estimators. We prove below that $\hat{\mu}$ and $\widehat{\mathfrak{C}}$ can tolerate, essentially, the smallest between $100\alpha\%$ and $100\beta\%$ of outlier contamination. We recommend using a large value of α, like $.50$ (α cannot exceed $.50$ by hypothesis in Proposition 2), but adjusting β according

to the expected proportion of outliers. In that case the breakdown point of the estimators will be essentially β.

Formally, the finite-sample breakdown point of $\hat{\mu}$, denoted by $\varepsilon_n^*(\hat{\mu})$, is defined as follows (Donoho and Huber 1983): given a sample $\mathcal{X} = \{X_1, \ldots, X_n\}$, let $\tilde{\mathcal{X}}$ be a contaminated sample obtained by replacing k points of $\mathcal{X}$ by arbitrary values; then $\varepsilon_n^*(\hat{\mu}) = k^*/n$, where k^* is the smallest value of k for which there is a sequence $\{\tilde{\mathcal{X}}^{(m)}\}$ of contaminated samples such that $\|\hat{\mu}^{(m)}\| \to \infty$. The asymptotic breakdown point $\varepsilon^*(\hat{\mu})$ is defined as the limit of $\varepsilon_n^*(\hat{\mu})$ when $n \to \infty$, provided the limit exists. The definitions of $\varepsilon_n^*(\widehat{\mathfrak{C}})$ and $\varepsilon^*(\widehat{\mathfrak{C}})$ are analogous.

Proposition 2 *Suppose $w(X_i) = g(\mathrm{rank}(r_i)/n)$, with g satisfying the conditions mentioned in Section 2. If $\alpha \leq .50$, $\lceil \alpha n \rceil \geq 3$ and $\beta \leq .50$, then $\varepsilon_n^*(\hat{\mu}) = \varepsilon_n^*(\widehat{\mathfrak{C}}) = \min(\lceil \alpha n \rceil, \lfloor \beta n \rfloor + 2)/n$ and $\varepsilon^*(\hat{\mu}) = \varepsilon^*(\widehat{\mathfrak{C}}) = \min(\alpha, \beta)$.*

The proof is given in the Appendix. In a few words, the proof shows that when the contaminating proportion is less than α, the $\lceil \alpha n \rceil$-th observation closest to an outlier is necessarily a non-outlier (regardless of the configuration of the data), whereas the $\lceil \alpha n \rceil$-th observation closest to a non-outlier is another non-outlier if the outliers are far enough from the bulk of the data. As a consequence, the radii of the outliers tend to be larger than the radii of the non-outliers, and trimming the largest radii will cut off the extreme outliers, as expected. This argument fails if there is a cluster of outliers with more than $\lceil \alpha n \rceil$ points, hence the importance of choosing a large α.

4 Functional forms and population parameters

Consider the family of probability distributions on $\mathcal{H}$, $\mathcal{P}$, which contains the empirical probability P_n of the random sample $X_1, \ldots, X_n$ as well as the population probability P from which the X_is are drawn. To study the consistency of the estimators it is instructive to derive their functional forms, that is, to find functionals $\mu : \mathcal{P} \to \mathcal{H}$ and $\mathfrak{C} : \mathcal{P} \to \mathcal{H}$ such that $\hat{\mu} = \mu(P_n)$ and $\widehat{\mathfrak{C}} = \mathfrak{C}(P_n)$, because under appropriate regularity conditions $\mu(P_n) \to \mu(P)$ and $\mathfrak{C}(P_n) \to \mathfrak{C}(P)$ in probability as $n \to \infty$ (Fernholz 1983; Van der Vaart 1998, chap. 20). In this paper we will

not be concerned with the precise conditions under which the estimators are consistent; instead, we will derive explicit expressions of $\mu(P)$ and $\mathfrak{C}(P)$ and see how they relate to the population mean and covariance of P, if they exist.

Given a probability P on $\mathcal{H}$ and a stochastic process X with distribution P, define $F_P(t; v) = P(\|X - v\| \leq t)$ for each $v \in \mathcal{H}$. The radius of the smallest ball centered at v that encompasses probability α is $r_P(v) = F_P^{-1}(\alpha; v)$, where $F_P^{-1}(\alpha; v) = \min\{t : F_P(t; v) \geq \alpha\}$, the usual quantile function. Then $r_P(X)$ is the α-radius around X, and if $G_P(t) := P\{r_P(X) \leq t\}$, the weight function $w_P(v)$ has the form $w_P(v) = g[G_P\{r_P(v)\}]$, with g as in Section 2. Then the functional forms of (2) and (3) are

$$\mu(P) = \frac{E_P\{w_P(X)X\}}{E_P\{w_P(X)\}}$$

and

$$\mathfrak{C}(P) = \frac{E_P[w_P(X)\{X - \mu(P)\} \otimes \{X - \mu(P)\}]}{E_P\{w_P(X)\}}.$$

The eigenvalues and eigenfunctions of $\mathfrak{C}(P)$ will be denoted by $\lambda_k(P)$ and $\phi_k(P)$, respectively.

The first question that arises is whether $\mu(P)$ and $\mathfrak{C}(P)$ are well defined for *any* $P \in \mathcal{P}$. The next proposition shows that this is indeed the case, because $r_P(v)$ bounds $\|v\|$ for any P and then all weighted moments of $\|X\|$ are finite even if $\|X\|$ itself does not have finite moments of any order.

Proposition 3 *For any $\alpha > 0$ and any $P \in \mathcal{P}$, there is a constant $K_{\alpha,P} \geq 0$ such that $\|v\| \leq r_P(v) + K_{\alpha,P}$ for all $v \in \mathcal{H}$. Therefore, if $\beta > 0$, $E_P\{w_P(X)\|X\|^k\} < \infty$ for any $k \geq 0$.*

Another important property of $r_P(v)$ is that it really is a measure of outlyingness, or depth, in the sense that $r_P(v)$ is smaller in regions of $\mathcal{H}$ where P is dense than in regions where P is sparse.

Proposition 4 *If v and w are two points in $\mathcal{H}$ such that $P(B_\delta(v)) \geq P(B_\delta(w))$ for all $\delta > 0$ (where $B_\delta(v)$ denotes the ball with center v and radius δ), then $r_P(v) \leq r_P(w)$.*

The equivariance of $\hat{\mu}$ and $\widehat{\mathfrak{C}}$ (Proposition 1) carries over to $\mu(P)$ and $\mathfrak{C}(P)$. That is, if P is the probability distribution of X and $\tilde{P}$ denotes the probability

distribution of $a\mathfrak{U}X + b$, with $\mathfrak{U}$ a unitary operator, $a \neq 0$, and $b \in \mathcal{H}$, then $r_{\tilde{P}}(v) = |a|\, r_P(v)$ for all $v \in \mathcal{H}$, $\mu(\tilde{P}) = a\mathfrak{U}\mu(P) + b$, and $\mathfrak{C}(\tilde{P}) = a^2\mathfrak{U}\mathfrak{C}(P)\mathfrak{U}^*$. The proof is given in the Appendix.

The next property shows that if the distribution of X has a point of symmetry μ_0, then $\mu(P) = \mu_0$. In that case $E(X)$, if finite, is also equal to μ_0, so the population mean and the trimmed mean coincide.

Proposition 5 *If X has a symmetric distribution about μ_0 (i.e. $X - \mu_0$ and $\mu_0 - X$ have the same distribution), then $\mu(P) = \mu_0$.*

Now suppose that X admits a decomposition of the form

$$X = \mu_0 + \sum_k \lambda_{0k}^{1/2} Z_k \phi_{0k} \tag{4}$$

with probability 1, where $\{Z_k\} \subset \mathbb{R}$ are random variables and $\mu_0 \in \mathcal{H}$, $\{\lambda_{0k}\} \subset \mathbb{R}$, and $\{\phi_{0k}\} \subset \mathcal{H}$ are parameters such that the ϕ_{0k}s are orthonormal and the λ_{0k}s are positive, non-increasing, and $\sum_k \lambda_{0k} < \infty$ (the sequence might be finite or infinite, so we omit the limits of summation). This decomposition holds, for instance, if X has finite second moments: in that case $\mu_0 = E(X)$, and $\{\lambda_{0k}\}$ and $\{\phi_{0k}\}$ are the eigenvalues and eigenfunctions of the covariance operator $E_P\{(X-\mu_0)\otimes(X-\mu_0)\}$ (this is a consequence of the spectral theorem; see Gohberg et al. 2003, chap. IV). In this case (4) is known as the Karhunen–Loève decomposition (Ash and Gardner 1975, chap. 1.4), and $Z_k = \langle X - \mu_0, \phi_{0k}\rangle / \lambda_{0k}^{1/2}$, which are uncorrelated with $E(Z_k) = 0$ and $V(Z_k) = 1$.

But expansion (4) is valid under more general conditions. For example, if the Z_ks are independent, Kolmogorov's Three Series Theorem (Gikhman and Skorokhod 2004, p. 384) implies that $\sum_k \lambda_{0k}^{1/2} Z_k \phi_{0k}$ converges almost surely in $\mathcal{H}$ if and only if $\sum_k P(\lambda_{0k} Z_{0k}^2 > c) < \infty$ for every $c > 0$. The latter is satisfied if the λ_{0k}s go to zero fast enough; it is not necessary that the Z_ks have finite moments. For instance, if each Z_k has a Cauchy distribution, then

$$\sum_k P(\lambda_{0k} Z_k^2 > c) \leq \sum_k \frac{2}{\pi} \left(\frac{\lambda_{0k}}{c}\right)^{1/2},$$

which converges for any $c > 0$ as long as $\sum_k \lambda_{0k}^{1/2} < \infty$.

Note that under (4) the interdistances $d_{ij} = \|X_i - X_j\|$ can be written as $d_{ij} = \{\sum_k \lambda_{0k}(Z_{ki} - Z_{kj})^2\}^{1/2}$. Therefore their distribution, and consequently that of the radii, will only depend on the Z_ks and the λ_{0k}s; the intrinsic elements of $\mathcal{H}$ in expansion (4), namely μ_0 and the ϕ_{0k}s, do not play any role in the distribution of the d_{ij}s. To put it in different words: the particular space $\mathcal{H}$ where the data lives does not determine the important properties of the estimators; the eigenvalues and the component scores do.

Proposition 6 *If expansion (4) holds with independent and symmetrically distributed Z_ks, then*

$$\mathfrak{C}(P) = \sum_k \tilde{\lambda}_{0k}(\phi_{0k} \otimes \phi_{0k})$$

for a positive sequence $\{\tilde{\lambda}_{0k}\}$ that does not depend on μ_0 or $\{\phi_{0k}\}$, but only on the distribution of $\{\lambda_{0k}^{1/2} Z_k\}$. In addition, if the Z_ks are identically distributed, then $\lambda_{0j} = \lambda_{0k}$ implies $\tilde{\lambda}_{0j} = \tilde{\lambda}_{0k}$.

Proposition 6 indicates that the set of trimmed principal components $\{\phi_k(P)\}$ coincides with the set $\{\phi_{0k}\}$, although we do not know whether the sequence $\{\tilde{\lambda}_{0k}\}$ is non-increasing, so it is not necessarily true that $\phi_k(P) = \phi_{0k}$ for each k. It is clear, however, that the dimensionality of the model is not changed, because

$$\tilde{\lambda}_{0k} = \frac{E_P\{w_P(X)|\langle X - \mu_0, \phi_{0k}\rangle|^2\}}{E_P\{w_P(X)\}}$$

and then $\{\tilde{\lambda}_{0k}\}$ has the same number of elements as $\{\lambda_{0k}\}$ if $\{\lambda_{0k}\}$ is a finite sequence. Moreover, the multiplicity of the eigenvalues associated with each ϕ_{0k} is preserved if the Z_ks are identically distributed.

5 Simulations

We ran some simulations to study the robustness and efficiency of the trimmed estimators, especially when compared to the spatial median and spherical principal

components (Gervini, 2008). One of the motivations for the present work was to improve on the low breakdown point of the spherical principal components, which we have accomplished, as shown below.

We simulated data from the model $X(t) = \sum_{k=1}^{\infty} Z_k \sqrt{\lambda_k} \phi_k(t)$ with Z_ks independent with $N(0,1)$ distribution, and $\phi_k(t) = \sqrt{2}\sin(\pi k t)$ for $t \in [0,1]$. Two sequences of eigenvalues were considered: a slow-decaying sequence $\lambda_k = 1/\{k(k+1)\}$ (Model 1) and a fast-decaying sequence $\lambda_k = 1/2^k$ (Model 2); note that $\sum_{k=1}^{\infty} \lambda_k = 1$ for both models. For practical purposes, Model 2 behaves like a finite-dimensional model, since the first five terms account for 97% of the total variability; Model 1, on the other hand, needs 31 terms to accumulate the same proportion of the variability, so it can be seen as a truly infinite-dimensional model. For the actual data generation we truncated Model 1 at the 1000th term and Model 2 at the 10th term, which represent 99.9% of the total variability in both cases. The curves were sampled on an equally spaced grid of 50 points. The sample size was $n = 100$ for all the scenarios described below, and each case was replicated 500 times.

First we studied the behavior of the sample mean, the median, trimmed means with hard-rejection weights, and trimmed means with soft-rejection weights (defined through the function (1)). For the hard-rejection trimmed means we considered the four possible combinations of the parameters $\alpha = .20$, $\alpha = .50$, $\beta = .20$, and $\beta = .50$. For the soft-rejection trimmed mean we also considered $\alpha = .20$ and $\alpha = .50$, but only $\beta = .20$ (the parameter β_1 was $.50$). In addition to the non-contaminated data, we generated outliers by adding $3\phi_1(t)$ to the first $n\varepsilon$ sample curves. Four contaminating proportions ε were considered: $.10$, $.20$, $.30$, and $.40$. The simulated root mean squared errors $\{E(\|\hat{\mu}\|^2)\}^{1/2}$ are reported in Table 1. In this table, and also in Table 2, "Hard(.20,.50)" denotes the hard-rejection trimmed estimator with $\alpha = .20$ and $\beta = .50$, and so on.

We see that the trimmed means compare favorably with the spatial median in terms of robustness. Although the median is more efficient for clean data, it becomes more biased at higher levels of contamination. Regarding the parameter α, we see that there are no advantages in taking $\alpha = .20$ instead of $\alpha = .50$. As for β, the hard-rejection trimmed mean with $\beta = .20$ is slightly more efficient than the hard-rejection trimmed mean with $\beta = .50$ for clean data, but much less robust at

	Model 1					Model 2				
	ε					ε				
Estimator	0	.10	.20	.30	.40	0	.10	.20	.30	.40
Mean	.10	.32	.61	.90	1.21	.10	.31	.61	.91	1.21
Median	.11	.17	.32	.54	.89	.11	.17	.32	.54	.88
Hard(.20,.20)	.13	.12	.14	.57	1.04	.13	.13	.15	.58	1.07
Hard(.50,.20)	.13	.12	.13	.30	.65	.12	.13	.13	.30	.66
Hard(.20,.50)	.16	.16	.16	.23	.78	.16	.16	.17	.23	.82
Hard(.50,.50)	.15	.15	.15	.16	.25	.15	.15	.15	.16	.30
Soft(.20,.20)	.14	.13	.13	.36	.93	.14	.14	.14	.37	.96
Soft(.50,.20)	.13	.13	.13	.17	.39	.13	.13	.14	.17	.44

Table 1: Simulation Results. Root mean squared errors of location estimators for different models and contamination proportions. Estimators are: sample mean, median, and trimmed means with hard- and soft-rejection weights.

high levels of contamination; so the choice $\beta = .50$ is clear in this case. The soft-rejection $(.50, .20)$ trimmed mean is slightly better than the hard-rejection $(.50, .50)$ trimmed mean for small εs, but the latter is more robust for large εs. So the choice between soft and hard trimming is not clear-cut from Table 1.

We also compared the behavior of estimators of the first principal component. We computed sample principal components, spherical principal components, and trimmed principal components with the same weights and parameters α and β as above. Now the outliers were generated in a different way: we added $3\phi_2(t)$ to the first $n\varepsilon/2$ sample curves and subtracted the same quantity to the next $n\varepsilon/2$ sample curves. This type of contamination inflates the variability in the direction of the second component but does not affect the mean. Therefore, a non-robust estimator of the first principal component will be closer to ϕ_2 than to ϕ_1, and the error $\|\hat{\phi}_1 - \phi_1\|$ will be approximately $\sqrt{2}$ (this would be considered breakdown for the first principal component). We considered the same four contamination proportions ε as before. The simulated root mean squared errors $\{E(\|\hat{\phi}_1 - \phi_1\|^2)\}^{1/2}$ are reported in Table 2.

	Model 1					Model 2				
	ε					ε				
Estimator	0	.10	.20	.30	.40	0	.10	.20	.30	.40
Sample p.c.	.13	1.34	1.38	1.39	1.39	.17	1.35	1.38	1.39	1.39
Spherical p.c.	.15	.31	1.00	1.32	1.37	.21	.61	1.20	1.33	1.37
Hard(.20,.20)	.22	.20	.15	1.32	1.38	.30	.30	.26	1.32	1.38
Hard(.50,.20)	.22	.21	.16	1.32	1.38	.31	.32	.26	1.32	1.38
Hard(.20,.50)	.35	.36	.33	.31	.32	.46	.47	.47	.46	.46
Hard(.50,.50)	.35	.37	.36	.34	.35	.43	.45	.49	.53	.68
Soft(.20,.20)	.27	.27	.25	.20	1.26	.37	.39	.37	.33	1.28
Soft(.50,.20)	.27	.28	.25	.20	.93	.35	.37	.39	.37	1.13

Table 2: Simulation Results. Root mean squared errors of first principal component estimators for different models and contamination proportions. Estimators are: sample p.c., spherical p.c., and trimmed p.c. with hard-rejection and soft-rejection weights.

The conclusions from Table 2 largely agree with those from Table 1, except that the soft-rejection $(.50, .20)$ estimator is now clearly preferable to the hard-rejection $(.50, .50)$ estimator; only for the extreme case $\varepsilon = .40$ is hard-rejection better. A noticeable fact in Table 2 is the low breakdown point of the spherical principal component: for all practical purposes, it breaks down at $\varepsilon = .20$, while the soft-rejection $(.50, .20)$ estimator does not seriously deteriorate until $\varepsilon = .40$, and the hard-rejection $(.50, .50)$ estimator has a low error even then. Although the trimmed principal components are considerably less efficient than the spherical principal components for non-contaminated data, we think that their superior robustness compensates for this.

Figure 2: Handwritten Digits Example. (a) Sample mean, (b) 41% trimmed mean, and (c) mean of the trimmed observations.

6 Applications

6.1 Handwritten Digits

This data was collected from 44 persons who wrote 250 samples of each digit $0, 1, \ldots, 9$. The dataset is available at the Machine Learning Repository of the University of California at Irvine, http://archive.ics.uci.edu/ml/, and has been previously analyzed by Izenman (2008, chap. 7). For each handwritten digit the trajectory of the pen's tip, $(x(t), y(t))$, was recorded at eight equally spaced time points. The data was rescaled and rotated so that x and y range between 0 and 100, and t between 0 and 1. Therefore, the sample space is $L^2([0,1]) \times L^2([0,1])$, which we endow with the canonical inner product $\langle (x_1, y_1), (x_2, y_2) \rangle = \int_0^1 x_1(t)x_2(t)dt + \int_0^1 y_1(t)y_2(t)dt$. We will only analyze the data corresponding to the digit "5", of which there were $n = 1055$ replications available.

At first glance, many of the handwritten digits in Figure 1 seem to correspond to the number "6" rather than "5". The reason is that $(x(t), y(t))$ does not track the vertical movement of the pen, and the way most people draw the number "5" is in two strokes: they begin at the upper left corner and move downwards, then raise the pen and draw the top dash. But there is another way to draw the number "5": in a single stroke, beginning at the upper right corner (like the letter "S"). The sample of 1055 handwritten "fives" is a mixture of both, and one soon realizes that

Figure 3: Handwritten Digits Example. Histogram of the radii with $\alpha = .50$.

the classical principal component analysis does not yield much useful information. The sample mean, shown in Figure 2(a), does not resemble a "5" or any other recognizable digit. When the leading 20 principal components are computed, one finds that the first component alone explains 78% of the variability. The scores of this component show a clear bimodal distribution, corresponding to the two types of "fives". Then the first principal component can be used for discrimination, but does not really provide useful information about subtler forms of variability in the written digits.

A better way to spot the two clusters is to use the radii. We computed radii for different values of α, and noticed that their distribution becomes increasingly bimodal as α increases. The histogram for $\alpha = .50$ is shown in Figure 3. There are two neatly distinguishable groups: 627 observations with $r_i < 60$, and 428 observations with $r_i > 60$. This does not necessarily imply that there are two clusters in the data, because large values of r_i that are close to each other do not necessarily correspond to curves that are close to each other in $\mathcal{H}$. But the large number of extreme radii (40.5% of the data) tends to point out to a systematic departure from a homogeneous population rather than isolated outliers. This is confirmed by a plot of the trimmed mean with $\beta = .41$ (Figure 2(b)), together with the mean of the observations that were cut off (Figure 2(c)). Clearly, Figure

Figure 4: Handwritten Digits Example. Effect of the trimmed principal components [(a) first, (b) second] on the trimmed mean. Solid line is the trimmed mean; dashed line is the trimmed mean plus 4 times the p.c.; dotted line is the trimmed mean minus 4 times the p.c.

2(b) (and then the observations with $r_i < 60$) correspond to the subclass of "fives" drawn in two strokes, whereas Figure 2(c) (and then the observations with $r_i > 60$) correspond to the more unusual "fives" drawn in one stroke.

The trimmed principal components do provide useful information about the main sources of variability in the bigger subclass. The first two components account for 56% and 14% of the total variability, respectively. The easiest way to interpret the components is to visualize their effects on the mean; so we see in Figure 4 that the first principal component mainly explains variation in the inclination of the "belly" of the digit, while the second principal component mostly explains variation in the slant of the vertical dash (which also has an effect on the position of the "belly").

The existence of more clusters can be investigated by recomputing the interdistances and the radii for each subclass independently. We did that and found that the radii had clear unimodal distributions, so there does not seem to be any other subclass of "fives" in addition to the two found in this section.

6.2 Excitation–Emission Matrices

In an enzyme cultivation process, real-time quality control is important because it allows for prompt adjustments in the process. To this end, samples of the cultivation broth are taken at regular intervals and the enzyme activity is measured. A traditional chemical analysis of the samples can determine enzyme activity very accurately, but it may take hours or days to perform. An alternative method employs multi-channel fluorescence sensors, which yield immediate results in the form of excitation-emission matrices (EEMs). But enzyme activity can be determined only indirectly from these matrices, using principal component analysis and partial least squares. The process is described in detail in Mortensen and Bro (2006); here we are going to use the data analyzed by these authors, which is available at http://www.models.life.ku.dk/research/data/.

Mortensen and Bro (2006) mention using a calibration sample of 283 EEMs, a test sample of 53 EEMs and an additional 15 EEMs of pure enzyme in their analysis. However, the calibration sample available online contains 338 EEMs (the test sample is also available but we do not use it here). It is not clear if Mortensen and Bro (2006) did a pre-screening of the calibration set and discarded some observations, but it is clear that the sample of 338 EEMs contains outliers. This is easy to see by plotting all the EEMs in quick succession; a movie showing this is available as supplementary material.

We carried out the principal component analysis on the logarithm of the light intensities, rather than directly on the light intensities. This ameliorates the effect of the outliers to some extent, but not completely. A histogram of the radii (Figure 5) shows that 15 observations are clear outliers (we suspect these are the pure-enzyme EEMs alluded to in the paper, although the website we downloaded the data from does not say so explicitly). There does not seem to be any other outliers, so we computed the hard-rejection trimmed mean and the leading 20 principal components with $\beta = .05$. The first component accounts for 59% of the total variability, and the second component for 20%. They are shown in Figure 6.

The sample mean and principal components are also shown in Figure 6. The first component explains 88% of the variability, and the second component only

Figure 5: Excitation–Emission Matrices Example. Histogram of the radii with $\alpha = .50$.

Figure 6: Excitation–Emission Matrices Example. (a) Trimmed mean, (b) first trimmed principal component, (c) second trimmed principal component, (d) sample mean, (e) first sample principal component, and (f) second sample principal component. Trimmed estimators were computed with 5% trimming.

Figure 7: Excitation–Emission Matrices Example. Scatter-plots of standardized component scores versus enzyme activity measure, for (a) the second trimmed principal component and (b) the second sample principal component.

5%. Note that while the sample mean and the trimmed mean do not differ very much, the first sample component is completely different from the first trimmed component. The reason is clear: as in the previous example, the first sample principal component only explains how the outliers vary from the non-outliers; it may be useful for outlier detection, but it is not associated with any genuine source of variability. Moreover, it downplays the importance of the second component, which is assigned only 5% of the total variability.

In fact, it is the second component the one primarily associated with enzyme activity. Figure 7 shows scatter plots of the standardized component scores versus an enzyme activity measure (the 15 outliers were removed). The association between component scores and enzyme activity, although not entirely linear, is clearly stronger for the trimmed principal component scores (the correlation coefficient is .69) than for the sample principal component scores (the correlation coefficient is .57).

These examples show that the principal components can be seriously affected by outliers in the data. These outliers can sometimes be detected by other means,

especially when they form a single cluster (for instance, by a classical discrimination method or by the first sample principal component, as in these examples). But isolated outliers are harder to detect because they have a masking effect on the estimators: they distort the principal components without appearing as unusual observations in a plot of the component scores or the residuals. Such cases can only be properly handled with robust estimators. A third example that illustrates this problem was not included for reasons of space but is available as supplementary material.

Appendix

Proof of Proposition 1

If $\tilde{X}_i = a\mathfrak{U}X_i + b$ with $\mathfrak{U}$ unitary, then $\tilde{d}_{ij} = \|(a\mathfrak{U}X_i + b) - (a\mathfrak{U}X_j + b)\| = |a|\,\|\mathfrak{U}(X_i - X_j)\| = |a|\,d_{ij}$. This implies that $\tilde{r}_i = |a|\,r_i$ and then $w(\tilde{X}_i) = w(X_i)$, because $\mathrm{rank}(\tilde{r}_i) = \mathrm{rank}(r_i)$. Therefore

$$\begin{aligned}\tilde{\mu} &= \frac{\sum_{i=1}^n w(\tilde{X}_i)\tilde{X}_i}{\sum_{i=1}^n w(\tilde{X}_i)} = \frac{\sum_{i=1}^n w(X_i)(a\mathfrak{U}X_i + b)}{\sum_{i=1}^n w(X_i)} \\ &= a\mathfrak{U}\hat{\mu} + b,\end{aligned}$$

and for $v \in \mathcal{H}$ we have

$$\begin{aligned}\widetilde{\mathfrak{C}}v &= \frac{1}{\sum_{i=1}^n w(\tilde{X}_i)} \sum_{i=1}^n w(\tilde{X}_i)\langle \tilde{X}_i - \tilde{\mu}, v\rangle(\tilde{X}_i - \tilde{\mu}) \\ &= \frac{1}{\sum_{i=1}^n w(X_i)} \sum_{i=1}^n w(X_i)\langle a\mathfrak{U}(X_i - \hat{\mu}), v\rangle\{a\mathfrak{U}(X_i - \hat{\mu})\} \\ &= a^2\mathfrak{U}\left\{\frac{1}{\sum_{i=1}^n w(X_i)} \sum_{i=1}^n w(X_i)\langle X_i - \hat{\mu}, \mathfrak{U}^*v\rangle(X_i - \hat{\mu})\right\} \\ &= a^2\mathfrak{U}\{\widehat{\mathfrak{C}}(\mathfrak{U}^*v)\},\end{aligned}$$

so $\widetilde{\mathfrak{C}} = a^2\mathfrak{U}\widehat{\mathfrak{C}}\mathfrak{U}^*$ as claimed.

Proof of Proposition 2

We will first show that $\varepsilon_n^*(\hat{\mu}) \geq \min(\lceil \alpha n \rceil, \lfloor \beta n \rfloor + 2)/n$, and then proceed to show that the equality holds by exhibiting a particular sequence of contaminations that makes $\|\hat{\mu}\|$ go to infinity.

Suppose, then, that $\{\tilde{\mathcal{X}}^{(m)}\}_{m \geq 1}$ is a sequence of contaminated samples obtained from $\mathcal{X}$ by replacing k observations, and such that $\|\hat{\mu}^{(m)}\| \to \infty$ as $m \to \infty$. Since $\|\hat{\mu}^{(m)}\| \leq \sum_{i=1}^n w(\tilde{X}_i^{(m)}) \|\tilde{X}_i^{(m)}\| / \sum_{i=1}^n w(\tilde{X}_i^{(m)})$, one can choose a sequence of points $\{\tilde{X}_{i_m}^{(m)}\}_{m \geq 1}$ such that $\|\tilde{X}_{i_m}^{(m)}\| \to \infty$ when $m \to \infty$ and $w(\tilde{X}_{i_m}^{(m)}) > 0$ for all m. The latter implies that the rank of the corresponding radii $\tilde{r}_{i_m}^{(m)}$ is strictly less than $(1-\beta)n$, because $g(t) = 0$ for any $t \geq 1 - \beta$; therefore $\tilde{r}_{i_m}^{(m)} < \tilde{r}_{(\lceil (1-\beta)n \rceil)}^{(m)}$.

Since $\tilde{r}_{i_m}^{(m)}$ is defined as the distance between $\tilde{X}_{i_m}^{(m)}$ and its $\lceil \alpha n \rceil$-th closest point in $\tilde{\mathcal{X}}^{(m)}$, and $\tilde{\mathcal{X}}^{(m)}$ has only k outliers, if $k < \lceil \alpha n \rceil$ there has to be at least one point X_{j_m} of the original sample $\mathcal{X}$ such that $\|\tilde{X}_{i_m}^{(m)} - X_{j_m}\| \leq r_{i_m}^{(m)}$. Since $\|\tilde{X}_{i_m}^{(m)}\| \to \infty$, it follows that $\|\tilde{X}_{i_m}^{(m)} - X_{j_m}\| \to \infty$ when $m \to \infty$ (the X_{j_m}s are bounded), so $\tilde{r}_{i_m}^{(m)} \to \infty$ as well. But $\tilde{r}_{i_m}^{(m)} < \tilde{r}_{(\lceil (1-\beta)n \rceil)}^{(m)}$, so a total of at least $1 + (n - \lceil (1-\beta)n \rceil + 1) = \lfloor \beta n \rfloor + 2$ radii $\tilde{r}_i^{(m)}$ go to infinity when $m \to \infty$.

Now, if $k < \lceil \alpha n \rceil$ then the number of non-outliers in $\tilde{\mathcal{X}}^{(m)}$ is $n - k \geq \lceil \alpha n \rceil$ (because $\alpha \leq .50$), so for each non-outlier X_i in $\tilde{\mathcal{X}}^{(m)}$ there are at least $\lceil \alpha n \rceil$ non-outliers X_j such that $\tilde{d}_{ij}^{(m)} = \|X_i - X_j\|$ remains bounded regardless of m; therefore the corresponding $\tilde{r}_i^{(m)}$s cannot go to infinity. This means that the $\lfloor \beta n \rfloor + 2$ radii that go to infinity correspond to observations in $\tilde{\mathcal{X}}^{(m)}$ that *are* outliers, so the number of outliers k cannot be less that $\lfloor \beta n \rfloor + 2$. Then $k \geq \lfloor \beta n \rfloor + 2$ when $k < \lceil \alpha n \rceil$. The other possibility is that $k \geq \lceil \alpha n \rceil$, so $k \geq \min(\lceil \alpha n \rceil, \lfloor \beta n \rfloor + 2)$. This proves that $\varepsilon_n^*(\hat{\mu}) \geq \min(\lceil \alpha n \rceil, \lfloor \beta n \rfloor + 2)/n$.

To see that $\varepsilon_n^*(\hat{\mu}) = \min(\lceil \alpha n \rceil, \lfloor \beta n \rfloor + 2)/n$, take $k = \min(\lceil \alpha n \rceil, \lfloor \beta n \rfloor + 2)$ and consider the two possibilities: $\lfloor \beta n \rfloor + 2 \leq \lceil \alpha n \rceil$ or $\lfloor \beta n \rfloor + 2 > \lceil \alpha n \rceil$. If $\lfloor \beta n \rfloor + 2 \leq \lceil \alpha n \rceil$, take any $X_0 \in \mathcal{H}$ with norm one and define the outliers $\tilde{X}_i^{(m)} = m^i X_0$, for $i = 1, \ldots, k$; then the distance between each $\tilde{X}_i^{(m)}$ and any other point in $\tilde{\mathcal{X}}^{(m)}$ (including other outliers) goes to infinity when $m \to \infty$, so $\tilde{r}_i^{(m)} \to \infty$ for $i = 1, \ldots, k$; since $k = \lfloor \beta n \rfloor + 2$ in this case, at least one of the outliers is not cut off and then $\|\hat{\mu}^{(m)}\| \to \infty$ as $m \to \infty$. For the other case, $\lfloor \beta n \rfloor + 2 > \lceil \alpha n \rceil$, define

the outliers $\tilde{X}_i^{(m)} = mX_0$ for $i = 1, \ldots, k$; then $\tilde{r}_i^{(m)} = 0$ for $i = 1, \ldots, k$ (because $k = \lceil \alpha n \rceil$ in this case), but $r_i^{(m)} > 0$ for the non-outliers (except in the trivial situation where all the sample points are identical). The number of non-outliers is $n - k = n - \lceil \alpha n \rceil > n - \lfloor \beta n \rfloor - 2$, so $n - k \geq n - \lfloor \beta n \rfloor - 1 \geq \lfloor \beta n \rfloor - 1$ (because $\beta \leq .50$). Therefore the $n - \lceil (1-\beta)n \rceil + 1 = \lfloor \beta n \rfloor + 1$ observations that are cut off include at most two outliers, and since $\lceil \alpha n \rceil \geq 3$ by hypothesis, there is at least one outlier that is not cut off, so $\|\hat{\mu}^{(m)}\| \to \infty$ as $m \to \infty$.

For $\varepsilon_n^*(\widehat{\mathfrak{C}})$ the proof is similar, because $\|\widehat{\mathfrak{C}}\| \leq \sum_{i=1}^n w(X_i)\|X_i\|^2 / \sum_{i=1}^n w(X_i) + \|\hat{\mu}\|$ and the preceding proof is also valid for $\sum_{i=1}^n w(X_i)\|X_i\|^2 / \sum_{i=1}^n w(X_i)$.

Proof of Proposition 3

By definition, $\alpha \leq P\{\|X - v\| \leq r_P(v)\}$ for any $v \in \mathcal{H}$. Since $\|X - v\| \geq |\|X\| - \|v\||$, it follows that $\alpha \leq P\{\|v\| - r_P(v) \leq \|X\|\}$. But if $\alpha > 0$, there is a finite $K_{\alpha,P}$ such that $P\{\|X\| > K_{\alpha,P}\} < \alpha/2$, say. Therefore $\|v\| - r_P(v) \leq K_{\alpha,P}$ for any $v \in \mathcal{H}$. Then, if $\beta > 0$, $w_P(v) > 0$ implies $r_P(v) \leq G_P^{-1}(1-\beta)$, which is finite, so

$$\begin{aligned} E_P\{w_P(X)\|X\|^k\} &\leq E_P[I\{\|X\| \leq K_{\alpha,P} + G_P^{-1}(1-\beta)\}\|X\|^k] \\ &\leq \{K_{\alpha,P} + G_P^{-1}(1-\beta)\}^k < \infty \end{aligned}$$

for any $k > 0$.

Proof of Proposition 4

Suppose $r_P(v) > r_P(w)$. By definition, $r_P(v) = \min\{\delta : P(B_\delta(v)) \geq \alpha\}$, so $P(B_{r_P(w)}(v)) < \alpha$. But $P(B_{r_P(w)}(v)) \geq P(B_{r_P(w)}(w))$ by hypothesis, and $P(B_{r_P(w)}(w)) \geq \alpha$ by definition of $r_P(w)$, a contradiction. Then it must be $r_P(v) \leq r_P(w)$.

Proof of the equivariance of $\mu(P)$ and $\mathfrak{C}(P)$

Let $X \sim P$ and $\tilde{X} = a\mathfrak{U}X + b \sim \tilde{P}$. Then

$$\begin{aligned} F_{\tilde{P}}(t; v) &= P(\|a\mathfrak{U}X + b - v\| \leq t) = P(\|X - \mathfrak{U}^*(v-b)/a\| \leq t/|a|) \\ &= F_P(t/|a|; \mathfrak{U}^*(v-b)/a), \end{aligned}$$

so $r_{\tilde{P}}(v) = |a| r_P(\mathfrak{U}^*(v-b)/a)$ for all $v \in \mathcal{H}$. Then

$$\begin{aligned} G_{\tilde{P}}(t) &= \tilde{P}\{r_{\tilde{P}}(\tilde{X}) \leq t\} = P\{r_{\tilde{P}}(a\mathfrak{U}X + b) \leq t\} \\ &= P\{|a| r_P(X) \leq t\} = G_P(t/|a|), \end{aligned}$$

which implies that

$$\begin{aligned} w_{\tilde{P}}(v) &= g[G_{\tilde{P}}\{r_{\tilde{P}}(v)\}] = g[G_P\{r_P(\mathfrak{U}^*(v-b)/a)\}] \\ &= w_P(\mathfrak{U}^*(v-b)/a) \end{aligned}$$

for all $v \in \mathcal{H}$. Therefore $E_{\tilde{P}}\{w_{\tilde{P}}(\tilde{X})\} = E_P\{w_{\tilde{P}}(a\mathfrak{U}X + b)\} = E_P\{w_P(X)\}$ and

$$E_{\tilde{P}}\{w_{\tilde{P}}(\tilde{X})\tilde{X}\} = E_P\{w_{\tilde{P}}(a\mathfrak{U}X + b)(a\mathfrak{U}X + b)\} = E_P\{w_P(X)(a\mathfrak{U}X + b)\},$$

from which it follows that $\mu(\tilde{P}) = a\mathfrak{U}\mu(P) + b$. In a similar way one obtains $\mathfrak{C}(\tilde{P}) = a^2\mathfrak{U}\mathfrak{C}(P)\mathfrak{U}^*$.

Proof of Proposition 5

It follows immediately from the equivariance of $\mu(P)$. Let $X \sim P$, $X - \mu_0 \sim P_1$ and $-X + \mu_0 \sim P_2$. Then $\mu(P_1) = \mu(P) - \mu_0$ and $\mu(P_2) = -\mu(P) + \mu_0$. But $P_1 = P_2$ by hypothesis, so $\mu(P) - \mu_0 = -\mu(P) + \mu_0$, which implies that $\mu(P) = \mu_0$.

Proof of Proposition 6

This result is also a direct consequence of the equivariance of the estimators. Since $\mathfrak{C}(\cdot)$ is location invariant we will assume, without loss of generality, that $\mu_0 = 0$.

Then, since the Z_ks have a symmetric distribution about 0, by Property 5 we have $\mu(P) = 0$. Therefore

$$\begin{aligned}
\mathfrak{C}(P) &= \frac{E_P\{w_P(X)(X \otimes X)\}}{E_P\{w_P(X)\}} \\
&= \frac{1}{E_P\{w_P(X)\}} \sum_j \sum_k E_P\{w_P(X)\lambda_{0j}^{1/2} Z_j \lambda_{0k}^{1/2} Z_k\}(\phi_{0j} \otimes \phi_{0k}) \\
&= \frac{1}{E_P\{w_P(X)\}} \sum_j \sum_k E_P\{w_P(X)\langle \phi_{0j}, X\rangle\langle \phi_{0k}, X\rangle\}(\phi_{0j} \otimes \phi_{0k}).
\end{aligned}$$

We will show that $E_P\{w_P(X)\langle \phi_{0j}, X\rangle\langle \phi_{0k}, X\rangle\} = 0$ for $j \neq k$. Since $\{\phi_{0k}\}$ is an orthonormal system in $\mathcal{H}$, if it is not complete we extend it to an orthonormal basis of $\mathcal{H}$, $\{\check{\phi}_{0k}\}$, and correspondingly we extend the sequences $\{\lambda_{0k}\}$ and $\{Z_k\}$ by adding zeros, so that we can write $X = \sum_k \check{\lambda}_{0k}^{1/2} \check{Z}_k \check{\phi}_{0k}$ with probability one. Now, any operator of the form $\mathfrak{S} = \sum_k a_k(\check{\phi}_{0k} \otimes \check{\phi}_{0k})$ with $a_k = \pm 1$ is unitary and self-adjoint, so we know by the proof of equivariance above that if $\tilde{P}$ denotes the probability distribution of $\mathfrak{S}X$, then $w_{\tilde{P}}(v) = w_P(\mathfrak{S}v)$ for all $v \in \mathcal{H}$. But under the current assumptions the $\check{Z}_k$s are independent and symmetrically distributed around 0 (including the artificially added $\check{Z}_k$s), so $\mathfrak{S}X$ and X have the same distribution; that is, $\tilde{P} = P$ and then $w_P(v) = w_P(\mathfrak{S}v)$ for all $v \in \mathcal{H}$. Therefore

$$\begin{aligned}
E_P\{w_P(X)\langle \check{\phi}_{0j}, X\rangle\langle \check{\phi}_{0k}, X\rangle\} &= E_P\{w_P(\mathfrak{S}X)\langle \check{\phi}_{0j}, \mathfrak{S}X\rangle\langle \check{\phi}_{0k}, \mathfrak{S}X\rangle\} \\
&= E_P\{w_P(X)\langle \mathfrak{S}\check{\phi}_{0j}, X\rangle\langle \mathfrak{S}\check{\phi}_{0k}, X\rangle\}.
\end{aligned}$$

If, in particular, we take $\mathfrak{S}_j$ to be the sign-change operator for the jth coordinate ($a_j = -1$ and $a_k = 1$ for any $k \neq j$), we have $\mathfrak{S}_j\check{\phi}_{0j} = -\check{\phi}_{0j}$ and $\mathfrak{S}_j\check{\phi}_{0k} = \check{\phi}_{0k}$ for any $k \neq j$, so

$$E_P\{w_P(X)\langle \mathfrak{S}_j\check{\phi}_{0j}, X\rangle\langle \mathfrak{S}_j\check{\phi}_{0k}, X\rangle\} = -E_P\{w_P(X)\langle \check{\phi}_{0j}, X\rangle\langle \check{\phi}_{0k}, X\rangle\}.$$

This implies that $E_P\{w_P(X)\langle \phi_{0j}, X\rangle\langle \phi_{0k}, X\rangle\} = 0$, as we wanted to prove. Then

$$\mathfrak{C}(P) = \sum_k \tilde{\lambda}_{0k}(\phi_{0k} \otimes \phi_{0k})$$

with

$$\tilde{\lambda}_{0k} = \frac{E_P\{w_P(X)|\langle\phi_{0k}, X\rangle|^2\}}{E_P\{w_P(X)\}},$$

as claimed.

Now, since $F_P(t; v) = P(\|X - v\| \leq t) = P\{\sum_k (\check{\lambda}_{0k}^{1/2}\check{Z}_k - \theta_k)^2 \leq t\}$, with $\theta_k = \langle\check{\phi}_{0k}, v\rangle$, it is clear that $r_P(v)$ depends on P only through $\{\lambda_{0k}^{1/2}Z_k\}$ and does not depend on μ_0 or on the ϕ_{0k}s. It is also clear that $r_P(v)$ depends on v only through its Fourier coefficients $\{\theta_k\}$, so the distribution of $r_P(X)$, $G_P(t)$, depends on P only through $\{\lambda_{0k}^{1/2}Z_k\}$ (because these are the Fourier coefficients of X). Consequently, the weight function $w_P(v) = g[G_P\{r_P(v)\}]$ depends on P only through $\{\lambda_{0k}^{1/2}Z_k\}$, and depends on v only through $\{\theta_k\}$, implying once again that the distribution of $w_P(X)$ depends only on $\{\lambda_{0k}^{1/2}Z_k\}$. Since $\langle\phi_{0k}, X\rangle = \lambda_{0k}^{1/2}Z_k$, it is clear then that $\tilde{\lambda}_{0k}$ depends on P only through $\{\lambda_{0k}^{1/2}Z_k\}$ and does not depend on μ_0 or the ϕ_{0k}s.

Finally, suppose that the Z_ks are identically distributed in addition to being independent, and that $\lambda_{0j} = \lambda_{0k}$. Then the distribution of X remains unchanged if we switch $\lambda_{0j}^{1/2}Z_j$ with $\lambda_{0k}^{1/2}Z_k$. More formally, define the "switch operator" $\mathfrak{S}_{jk} = (\phi_{0j} \otimes \phi_{0k} + \phi_{0k} \otimes \phi_{0j}) + \{\mathfrak{I} - (\phi_{0j} \otimes \phi_{0k} + \phi_{0k} \otimes \phi_{0j})\}$, where $\mathfrak{I}$ is the identity operator. Then, if $X = \sum_k \lambda_{0k}^{1/2}Z_k\phi_{0k}$, we have $\mathfrak{S}_{jk}X = \lambda_{0j}^{1/2}Z_j\phi_{0k} + \lambda_{0k}^{1/2}Z_k\phi_{0j} + \sum_{l\neq j,k}\lambda_{0l}^{1/2}Z_l\phi_{0l}$, so X and $\mathfrak{S}_{jk}X$ have the same distribution. Moreover, since $\mathfrak{S}_{jk}$ is unitary and self-adjoint, $w_P(v) = w_P(\mathfrak{S}v)$ for all $v \in \mathcal{H}$, as before, so

$$\begin{aligned} E_P\{w_P(X)|\langle\phi_{0j}, X\rangle|^2\} &= E_P\{w_P(\mathfrak{S}_{jk}X)|\langle\phi_{0j}, \mathfrak{S}_{jk}X\rangle|^2\} \\ &= E_P\{w_P(X)|\langle\mathfrak{S}_{jk}\phi_{0j}, X\rangle|^2\} \\ &= E_P\{w_P(X)|\langle\phi_{0k}, X\rangle|^2\}. \end{aligned}$$

Then $\tilde{\lambda}_{0j} = \tilde{\lambda}_{0k}$, as claimed.

References

Ash, R. B., and Gardner, M. F. (1975), *Topics in Stochastic Processes*, Probability and Mathematical Statistics (Vol. 27), New York: Academic Press.

Cuevas, A., Febrero, M., and Fraiman, R. (2007), "Robust Estimation and Classi-

fication for Functional Data via Projection-Based Depth Notions," *Computational Statistics*, 22, 481–496.

Cuevas, A., and Fraiman, R. (2009), "On Depth Measures and Dual Statistics. A Methodology for Dealing with General Data," *Journal of Multivariate Analysis*, 100, 753–766.

Davies, P. L. (1987), "Asymptotic Behavior of S-estimates of Multivariate Location Parameters and Dispersion Matrices," *The Annals of Statistics*, 15, 1269–1292.

Donoho, D., and Huber, P. (1983), "The Notion of Breakdown Point," in *A Festschrift for Erich L. Lehmann*, P. J. Bickel, K. A. Doksum and J. L. Hodges (eds.), Belmont, CA: Wadsworth, pp. 157–184.

Fernholz, L. T. (1983), *Von Mises Calculus for Statistical Functionals*, Lecture Notes in Statistics No. 19, New York: Springer.

Filzmoser, P., Maronna, R., and Werner, M. (2008), "Outlier Identification in High Dimensions," *Computational Statistics and Data Analysis*, 52, 1694–1711.

Filzmoser, P., Serneels, S., Maronna, R., and Van Espen, P. J. (2009), "Robust Multivariate Methods in Chemometrics," in *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis* (Vol. III), Amsterdam: Elsevier, pp. 681–722.

Fraiman, R., and Muniz, G. (2001), "Trimmed Means for Functional Data," *Test*, 10, 419–40.

Gervini, D. (2008), "Robust Functional Estimation Using the Spatial Median and Spherical Principal Components," *Biometrika*, 95, 587–600.

Gikhman, I. I., and Skorokhod, A. V. (2004), *The Theory of Stochastic Processes I*, New York: Springer.

Gohberg, I., Goldberg, S., and Kaashoek, M. A. (2003), *Basic Classes of Linear Operators,* Basel: Birkhäuser Verlag.

Hsieh, A. D., and Hung, Y. S. (2009), "Detecting Outlier Samples in Microarray Data," *Statistical Applications in Genetics and Molecular Biology*, 8 (1), article number 13.

Izenman, A. J. (2008), *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, Springer Texts in Statistics, New York: Springer.

Jolliffe, I. T. (2002), *Principal Component Analysis* (2nd ed.), Springer Series in Statistics, New York: Springer.

Liu, R. Y. (1990), "On a Notion of Data Depth Based on Random Simplices," *The Annals of Statistics*, 18, 405–414.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., and Cohen, K. L. (1999), "Robust Principal Component Analysis for Functional Data" (with discussion), *Test*, 8, 1–73.

López-Pintado, S., and Romo, J. (2009), "On the Concept of Depth for Functional Data," *Journal of the American Statistical Association*, 104, 718–734.

Maronna, R. A. (1976), "Robust M-estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51–67.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006), *Robust Statistics. Theory and Methods*, Wiley Series in Probability and Statistics, New York: Wiley.

Mortensen, P. P., and Bro, R. (2006), "Real-Time Monitoring and Chemical Profiling of a Cultivation Process," *Chemometrics and Intelligent Laboratory Systems*, 84, 106–113.

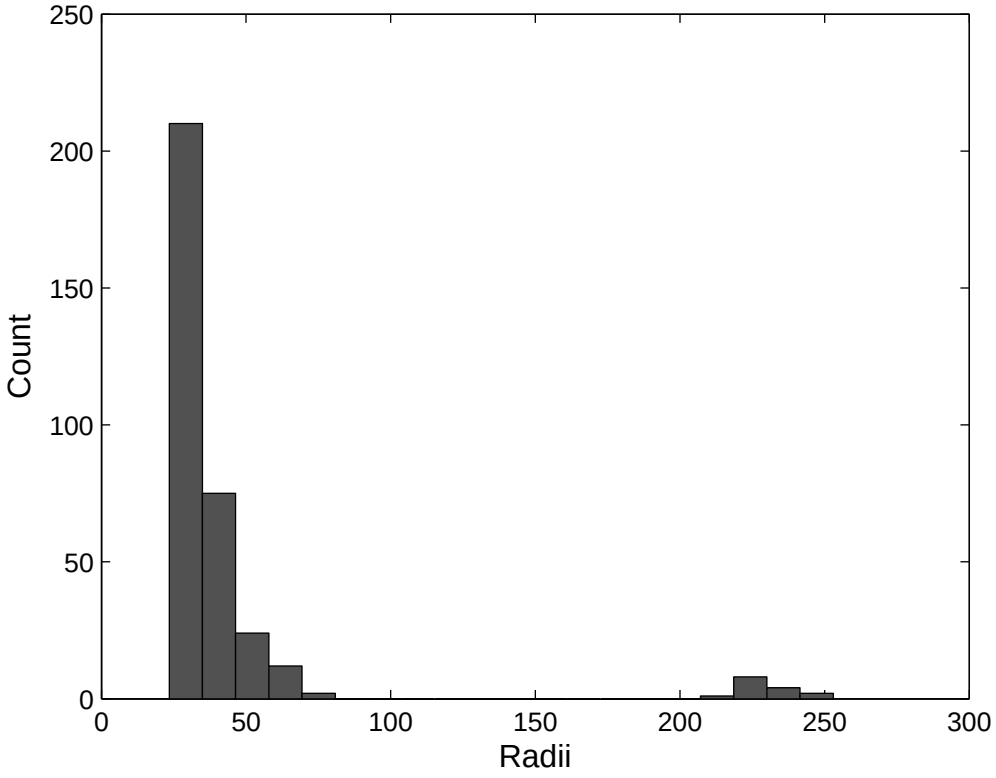
Oja, H. (1983), "Descriptive Statistics for Multivariate Distributions," *Statistics and Probability Letters*, 1, 327–332.

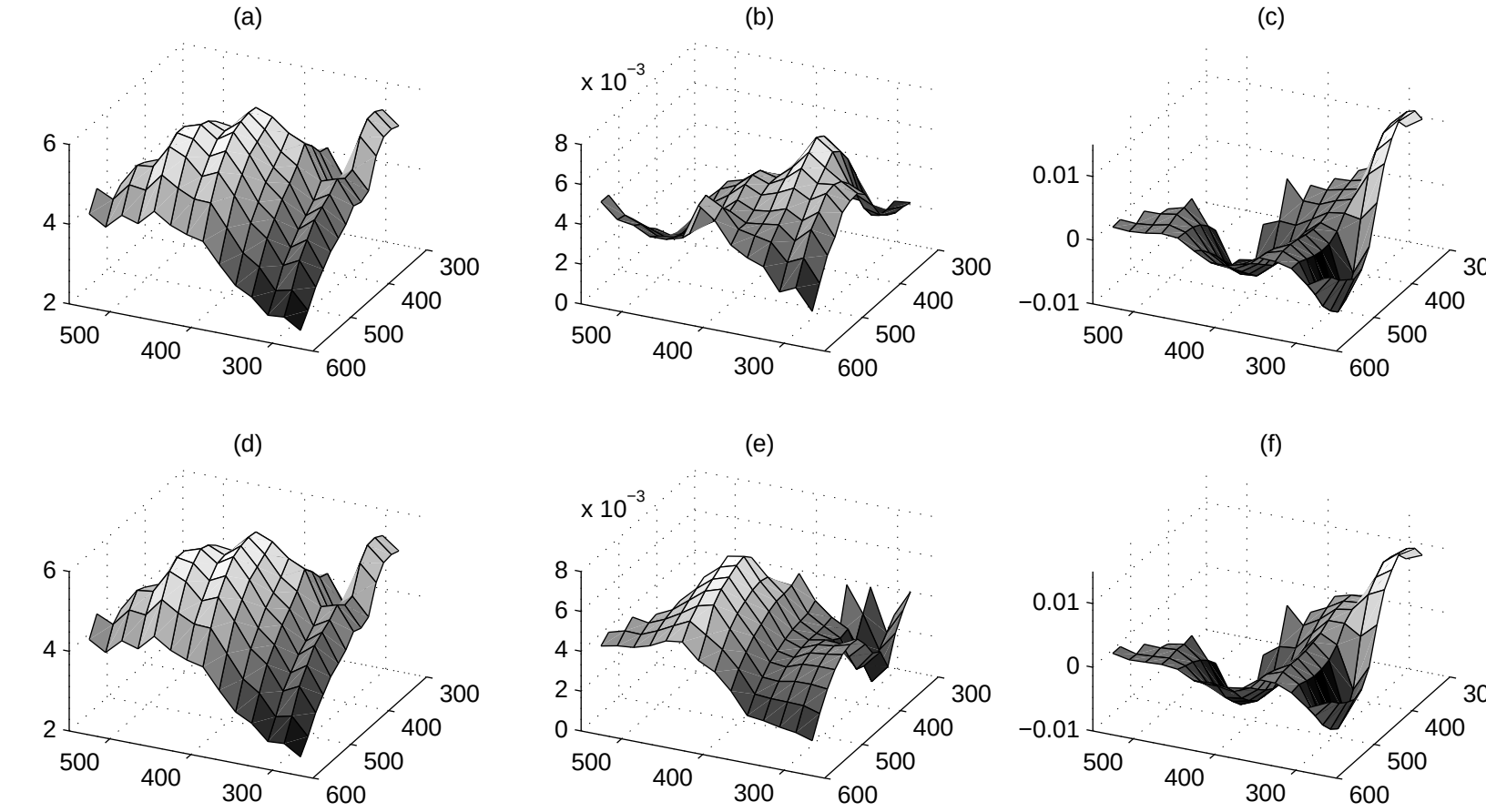
Ramsay, J. O., and Silverman, B. W. (2002), *Applied Functional Data Analysis. Methods and Case Studies*, Springer Series in Statistics, New York: Springer.

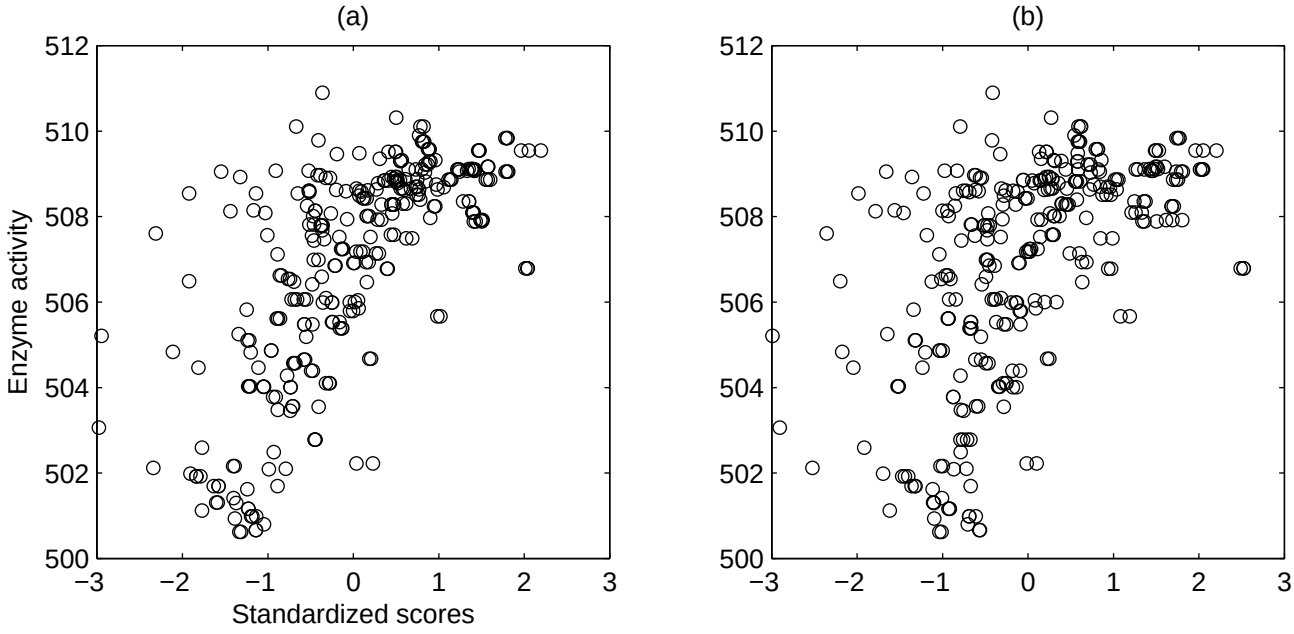
Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis* (2nd ed.), Springer Series in Statistics, New York: Springer.

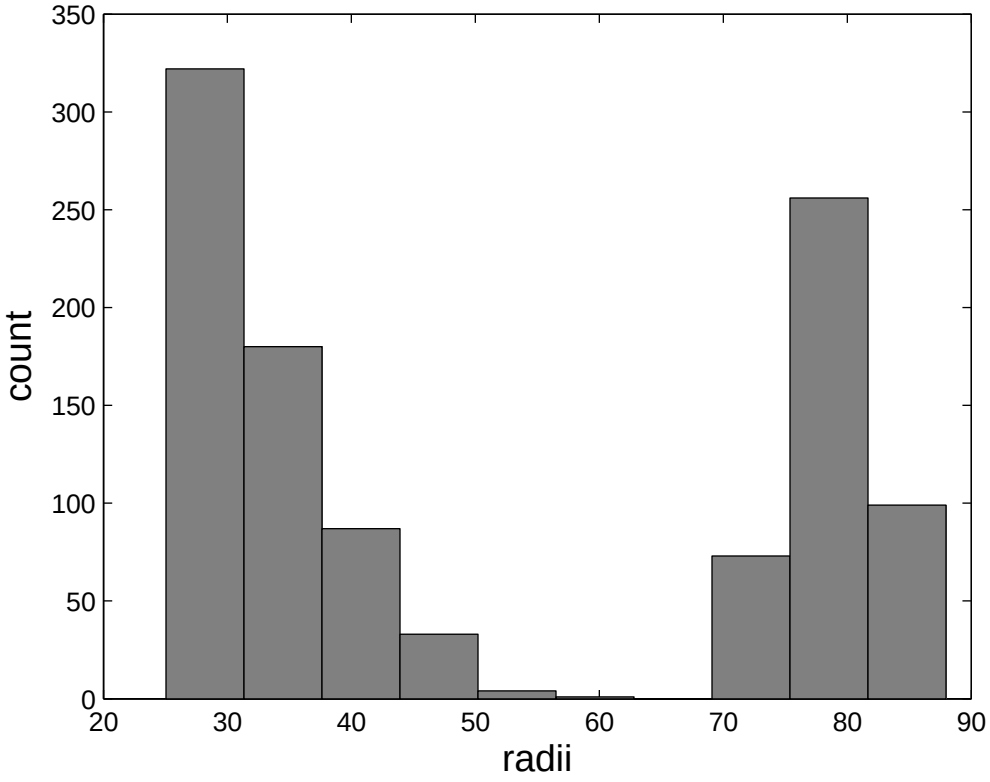
Van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge, UK: Cambridge University Press.

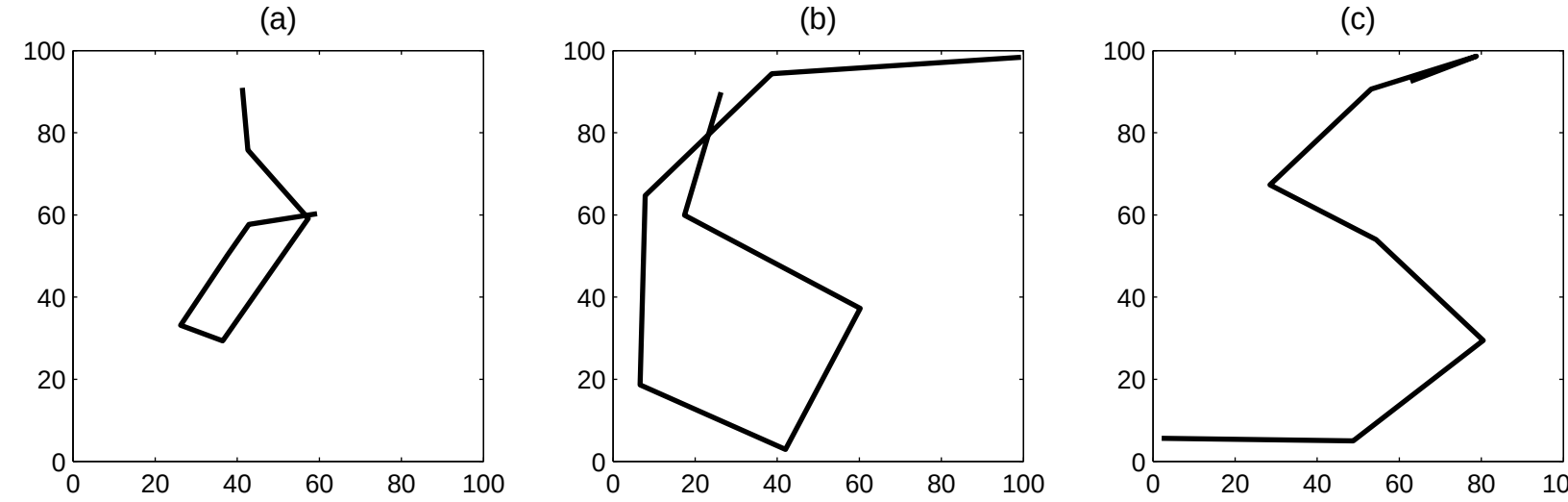
Wang, N., and Raftery, A. E. (2002), "Nearest-Neighbor Variance Estimation (NNVE): Robust Covariance Estimation via Nearest-Neighbor Cleaning" (with discussion), *Journal of the American Statistical Association*, 97, 994–1019.

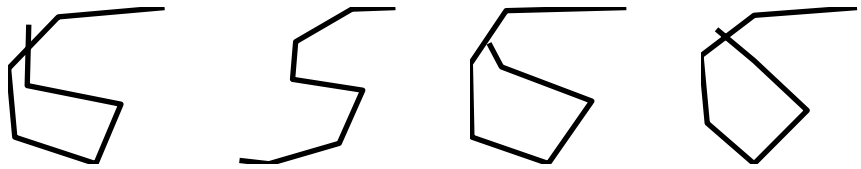

250
200
150
100
50
0
Count
0
50
100
150
200
250
300
Radii

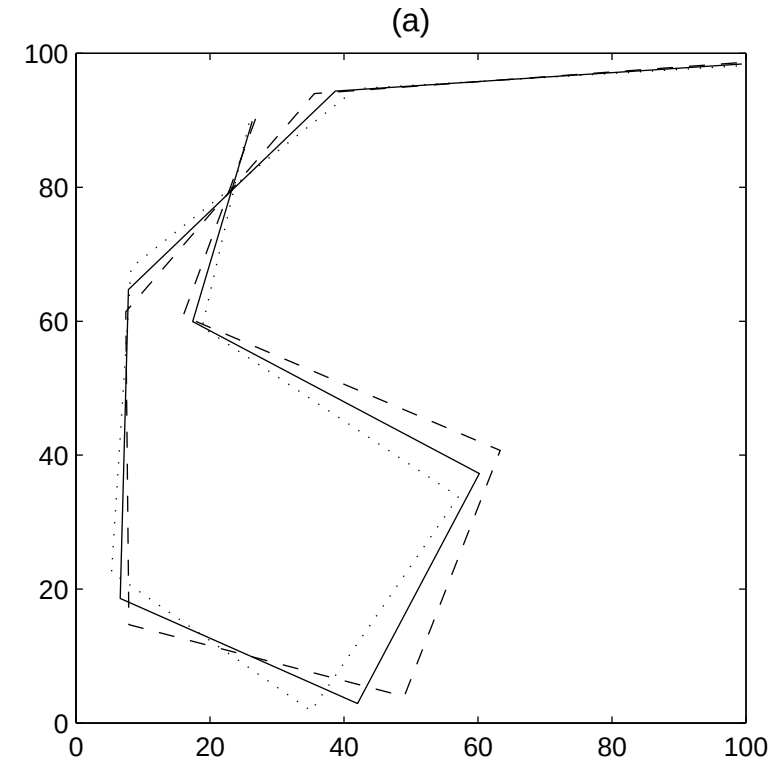

(a)
(b)
(c)
x 10^−3
6
4
2
8
0
0.01
−0.01
500
400
300
600
(d)
(e)
(f)

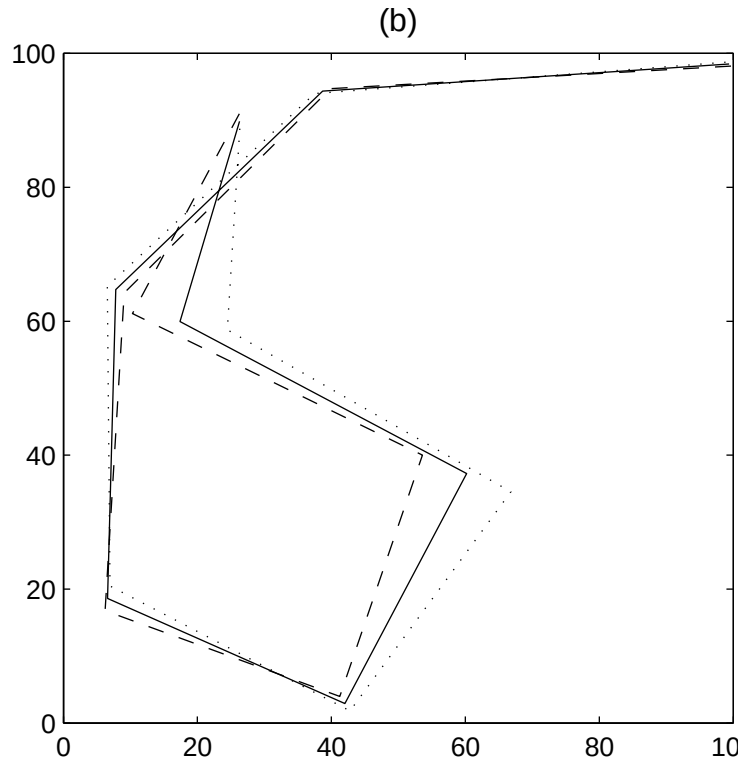

(a)
(b)
Enzyme activity
Standardized scores
512
510
508
506
504
502
500
−3
−2
−1
0
1
2
3


350
300
250
200
150
100
50
0
count
20
30
40
50
60
70
80
90
radii


(a)
(b)
(c)
0
20
40
60
80
100




(a)
100
80
60
40
20
0
0
20
40
60
80
100


(b)
100
80
60
40
20
0
0
20
40
60
80